Cowell, R., Graversen, T., Lauritzen, S. L. & Mortera, J. (2014). Analysis of forensic DNA mixtures with artefacts. Journal of the Royal Statistical Society. Series C: Applied Statistics, doi: 10.1111/rssc.12071

Analysis of Forensic DNA Mixtures with Artefacts

R. G. Cowell

City University London, United Kingdom

T. Graversen

University of Oxford, United Kingdom

S. L. Lauritzen

University of Oxford, United Kingdom

J. Mortera

Università Roma Tre, Italy

Summary. DNA is now routinely used in criminal investigations and court cases, although DNA samples taken at crime scenes are of varying quality and therefore present challenging problems for their interpretation. We present a statistical model for the quantitative peak information obtained from an electropherogram (EPG) of a forensic DNA sample and illustrate its potential use for the analysis of criminal cases. In contrast to most previously used methods, we directly model the peak height information and incorporate important artefacts associated with the production of the EPG. Our model has a number of unknown parameters, and we show that these can be estimated by the method of maximum likelihood in the presence of multiple unknown individuals contributing to the sample, and their approximate standard errors calculated; the computations exploit a Bayesian network representation of the model. A case example from a UK trial, as reported in the literature, is used to illustrate the efficacy and use of the model, both in finding likelihood ratios to quantify the strength of evidence, and in the deconvolution of mixtures for the purpose of finding likely profiles of the individuals contributing to the sample. Our model is readily extended to simultaneous analysis of more than one mixture as illustrated in a case example. We show that combination of evidence from several samples may give an evidential strength close to that of a single source trace and thus modelling of peak height information provides for a potentially very efficient mixture analysis.

Keywords: allelic dropout; Bayesian networks; DNA profiles; forensic statistics; gamma distribution; mixture deconvolution; strength of evidence; stutter.

1. Introduction

Since the pioneering work of Jeffreys et al. (1985), genetic fingerprinting or DNA profiling has become an indispensable tool for identification of individuals in the investigative and judicial process associated with criminal cases, in paternity and immigration cases, and other contexts (Aitken and Taroni, 2004; Balding, 2005). The technology has now reached a stage where uncertainty associated with determining the identity of an individual based on matching a high quality DNA sample from a crime scene to one taken under laboratory conditions has been virtually eliminated. Frequently, however, DNA samples found on

Address for correspondence: Steffen L. Lauritzen, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.
E-mail: steffen@stats.ox.ac.uk

crime scenes are more complex, either because the amounts of DNA are tiny, even just a few molecules, the samples contain DNA from several individuals, the DNA molecules have degraded, or a combination of these. In such cases there is considerable uncertainty involved in determining whether or not the DNA of a given individual, for example a suspect in a criminal case, is actually present in the given sample. A sample found on the crime scene may contain material from the victim but also from individuals which might be involved in the crime, and determining their DNA becomes an important issue for the criminal investigation. In many cases such *mixed* samples contain DNA from, say, a victim and a perpetrator, but advanced DNA technology now extracts genetic material from a huge variety of surfaces and objects which may have been handled by several individuals, only some of them being related to the specific crime. DNA mixtures with DNA from many individuals occur frequently in multiple rape cases or in traces left by groups of perpetrators handling the same objects such as, for example, balaclavas and crowbars.

The basic representation of the composition of a DNA sample is the *electropherogram* (EPG), as described further in Section 1.1 below and displayed in Fig. 1. The identification of the DNA composition from the EPG information gives rise to a wide range of challenging statistical questions, some associated with uncertainties and artefacts in the measurement processes, some associated with population genetic variations, and other of a more conceptual nature. This article attempts to address some of these challenges.

As detailed in Section 3 below, the statistical issues concerning DNA mixtures have previously been addressed in a number of different ways. Some approaches make use only of the presence or absence of peaks in the EPG or use the peak heights in a semi-quantitative fashion; other approaches seek to develop fully quantitative models of observed peak heights. Although several methods have been developed, they each suffer from limitations of various kinds and none of the methods have been universally acknowledged as gold standards.

This article presents a new statistical model for the peak heights of an EPG. It builds upon earlier models presented in Cowell et al. (2007a,b, 2011) but is simpler than these, for example eliminating the need for previously introduced discretizations of some continuous parameters. The simplifications of the model combined with an efficient Bayesian network representation (Graversen and Lauritzen, 2013a) enables fast computation and permits analysis of mixtures in which the presence of several unknown contributors is posited; in particular, estimation of unknown parameters by maximum likelihood becomes computationally feasible.

The plan of the paper is as follows. In the next section we present background information concerning DNA relevant for mixture analyses carried out by forensic scientists. We summarize the measurement processes carried out for quantifying DNA mixtures, and the artefacts that can arise in these processes and lead to difficulties in their interpretation. A case example from a trial described in the literature is presented in Section 1.2 and used for illustration in the remainder of the paper. In Section 1.3 we describe the basic elements associated with statistical interpretation of DNA evidence. In Section 2 we give a detailed description of our basic model for peak heights and its elaboration with inclusion of various artefacts associated with the EPG and we briefly review related work on DNA mixture modelling in Section 3; we note that also ?) gives a review of recent literature and software for DNA mixture analysis. In Section 4 we apply our model to the case presented in Section 1.2, estimating the unknown parameters in our model by maximum likelihood using an efficient Bayesian network representation. Section 4.3 is devoted to showing how the modularity of the Bayesian network representation may be used to elicit further details in the analysis of a mixture, for example by finding the probability that a particular peak corresponds to a proper allele or is due to an *artefact* of the measurement process such as, for example, stutter; see further in Section 1.1.2. It is also shown how the networks may be extended for other purposes. In Section 4.4 we show how the model can also be useful for the analysis of DNA from a single individual. Finally we discuss suggestions for further work in Section 5.

1.1. DNA mixtures

In this section we describe the nature of the information that we analyse with our DNA mixture model and give a brief description of the DNA amplification process and associated measurements as carried out in forensic laboratories. We also introduce some of the complications that can occur. A more comprehensive description may be found in Butler (2005).

1.1.1. *Short Tandem Repeat (STR) markers*

Forensic scientists encode an individual's genetic profile using the composition of DNA at various positions on the chromosomes.

A specific position on a chromosome is called a *locus*, or *marker*. The information at each locus consists of an unordered pair of *alleles* which forms the *genotype* at that locus; a pair because chromosomes come in pairs, one inherited from the father and one from the mother, and unordered because it is not recorded from which chromosome of the pair each allele originates. We adopt the standard assumption that the population is in Hardy–Weinberg equilibrium implying that the two alleles of an individual can be assumed to be sampled independently from the same population. Human DNA has twenty three pairs of chromosomes: twenty two autosomal chromosome pairs and a sex-linked pair.

The loci used for forensic identification have been chosen for various reasons. Among these, we point out the following two. The first reason is that at each locus there is a wide variability between individuals in the alleles that may be observed. This variability can therefore be exploited to differentiate people. The second reason is that each locus is either on a distinct chromosome, or if a pair of loci are on the same chromosome then they are widely separated. This means that the alleles at the various loci may be treated as mutually independent, thus simplifying the statistical analysis.

The alleles of a marker are sequences of the four amino acid nucleotides *adenine, cytosine, guanine* and *thymine*, which we represent by the letters A, C, G and T. Each amino acid is also called a *base*, and because the DNA molecule has a double helix structure, each amino acid on one strand is linked to a complementary amino acid on the other strand; a complementary pair of amino acids is called a *base pair*.

An allele is typically named by its *repeat number*, usually an integer. For example, consider the allele with repeat number 5 (commonly also referred to as allele 5 for brevity) of the marker TH01. This allele includes the sequence of four nucleotides AATG repeated consecutively five times. It can be designated by the formula $[AATG]_5$. Likewise allele 8 of TH01 has eight consecutive repetitions of the AATG sequence, which may be denoted by $[AATG]_8$. Repeat numbers are not always integers. For example, allele 8.3 of TH01 has the chemical sequence $[AATG]_5ATG[AATG]_3$, in which '8' refers to the eight complete four-word bases $[AATG]$ and the '.3' refers to the three base-long word sequence ATG in the middle. Repeat numbers with decimal '.1' and '.2' endings are also possible, indicating the presence of a word of one or two bases. Note that the integer part of the repeat number counts how many complete words of four bases make up the allele sequence but the words need not be all identical and may vary even within loci. For example, allele 11 of the marker VWA has the base sequence $TCTA[TCTG]_3[TCTA]_7$. Also, some markers are based on tri- or pentanucleotide motifs rather than tetranucleotides as above. The base-letter sequences for many alleles may be found in Butler (2005).

When the repeat numbers of the two alleles of an individual at a marker are the same, then the genotype for that marker is said to be *homozygous*; when the repeat numbers differ, the genotype for that marker is said to be *heterozygous*.

The repetitive structure in the alleles gives rise to the term *short tandem repeat* (STR) marker to describe these loci; they also go by the name of *microsatellites*. Note that for other purposes of genetic analysis it is common to use single-nucleotide polymorphisms (SNPs) which are defined as DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. However, for a number of reasons, the use of STR alleles is dominating in forensic genetics.

Within a population the various alleles of STR markers do not occur equally often, some can be quite common and some quite rare. When carrying out probability calculations based on DNA, forensic scientists use estimates of probabilities based on allele frequencies in profiles of a sample of individuals. The sample sizes typically range from a few hundred to thousands of individuals. For example, Butler et al. (2003) presents tables of US-population STR-allele frequencies for Caucasians, African-Americans, and Hispanics based on sample sizes of 302, 258 and 140 individuals.

1.1.2. The PCR process

The DNA collected from a crime scene for forensic analysis consists of a number of human cells from one or more individuals. Note that each cell of an individual will contain two alleles (diploid cells) for each autosomal marker, whereas sperm cells have only one allele (haploid cells). This means that in a mixture, a particular individual will contribute the same number of alleles for each marker. In order to identify the alleles that are present, a DNA sample is first subjected to chemical reagents that break down the cell walls so that the individual chromosomes are released into a solution. A small amount of this solution is exploited to quantify the concentration of DNA; the typical unit of measurement is picograms per microlitre, the DNA in a single human cell having a mass of around six picograms. Having determined the density of DNA in the sample, a volume is extracted that is estimated to contain a certain quantity of DNA, typically around 0.5 nanograms, equivalent to around 100 human cells. The DNA in the extract is then amplified using the *polymerase chain reaction* (PCR) process. This involves adding *primers* and other biochemicals to the extract, and then subjecting it to a number of rapid heating and cooling cycles. Heating the extract has the effect of splitting apart the two complementary strands of DNA, the cooling phase then allows free floating amino acids to bind with these individual strands in such a way that the DNA is copied. By the action of repeated heating and cooling cycles, typically around 28 altogether, an initially small amount of DNA is amplified to an amount large enough for quantification. Mathematically, the amplification process may be modelled as a branching process (Sun, 1995; Stolovitzky and Cecchi, 1996). The amplification process is not 100% efficient, that is, not every allele gets copied in each cycle. This means that if two distinct alleles in a marker are present in the extract in the same amount prior to amplification, they will occur in different amounts at the end of the PCR process. In our model we capture this variability using gamma distributions.

Note that after breaking down the cell walls to release the chromosomes there could be sufficient DNA in the sample to carry out PCR amplifications with several extracts. When this is done it is called a *replicate run.*

To understand the quantification stage of the post PCR amplified DNA, it is important to know that the amplification process does not copy only the repeated DNA word segment of a marker, it also copies DNA at either end. These are called *flanking regions*, and their presence is important in performing the PCR process. Thus an amplified allele will consist of the allele word sequence itself and two flanking regions, and will have a length associated with it which is measured in the total number of base pairs included in the word sequence and the flanking regions. For each marker, the DNA sequence (hence the size) forming each of the two flanking regions is constant, but different across markers. Thus quantifying a certain allele is equivalent to measuring how much DNA is present of a certain size. This is carried out by the process of *electrophoresis*, as follows.

The flanking regions have attached to them a fluorescent dye. Several colours of fluorescent dyes are used to distinguish similarly sized alleles from different markers. The amplified DNA is drawn up electrostatically through a fine capillary to pass through a light detector, which illuminates the DNA with a laser and measures the amount of fluorescence generated. The latter is then an indication of the number of alleles tagged with the fluorescent marker. The longer alleles are drawn up more slowly than the shorter alleles, however alleles of the same length are drawn up together. This means that the intensity of the detected fluorescence will sharply peak as a group of alleles of the same length passes the light detector, and the value of the intensity will be a measure of the number of alleles that pass. The detecting apparatus thus measures a time series of fluorescent intensity, but it converts the time variable into an equivalent base pair length variable. The data may be presented to a forensic scientist as an *electropherogram* (EPG) as shown in Fig. 1 with each panel in the EPG corresponding to a different dye. The horizontal axes indicate the base pair length, and the vertical axis the intensity.

In the absence of artefacts, a peak in the EPG indicates presence of an allele in the sample before amplification. The peak height is a measure of the amount of the allele in the amplified sample expressed in *relative fluorescence units* (RFU). The area of the peak is another measure of the amount, but this is highly correlated with the height (Tvedebrink et al., 2010). Both peak height and peak area are determined by software in the detecting apparatus.

We shall call the peak size information extracted from the EPG the *profile of the DNA sample*, or more briefly, the DNA profile. Commonly, DNA profile also refers to the combined genotype of a person across all markers.

In measuring the peak heights, low level noise give rise to small spurious peaks. A *peak amplitude threshold* may be set by the forensic analyst whereby peaks below the threshold level are ignored. Thus an allele present in the DNA sample will not be recorded as observed if the peak it generates is below the threshold; when this happens a *dropout* of the allele is said to have occurred. A dropout can also occur simply by sporadic failure of the apparatus. Dropout is an artefact that can make the analysis of DNA samples difficult. Another common artefact is *stutter*, whereby an allele that is present in the sample is mis-copied at some stage in the PCR amplification process, and a four base pair word segment is omitted; more rarely other stutter patterns occur. This damaged copy itself takes part in the amplification process, and so yields a peak four base pairs below the allele from which it arose.

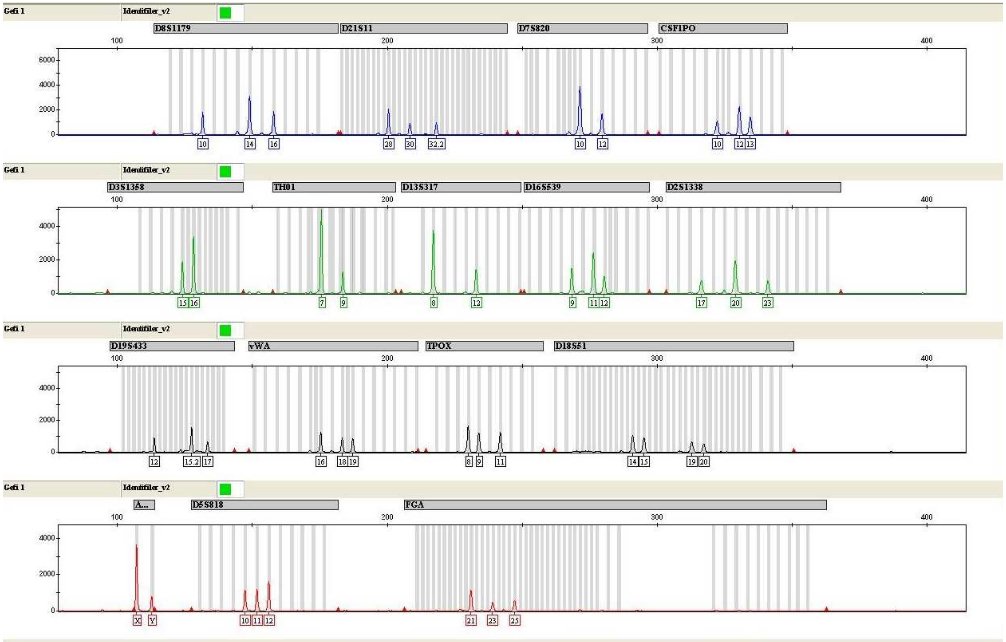

Fig. 1. An electropherogram (EPG) using the Identifiler$^{\text{TM}}$ STR kit. Each of the four panels represents a dye. Marker names are in grey boxes above panels and repeat numbers of detected alleles are given below the corresponding peaks. The vertical grey bars are used to identify alleles by their length in base pairs. On the left-hand side of each panel the scale of RFU is indicated.

Another artefact is known as *dropin*, referring to the occurrence of small unexpected peaks in the EPG. This can for example be due to sporadic contamination of a sample either at source or in the forensic laboratory.

Current technology allows for the amplification of very little DNA material, even as little as contained within one cell (Findlay et al., 1997). Amplifications of low amounts of DNA are termed low template DNA (LTDNA). LTDNA analyses are particularly prone to artefacts such as dropin and dropout.

Finally, a mutation in the flanking region can result in the allele not being picked up at all by the PCR process, in which case we say that the allele is *silent*. An allele can also be undetectable and thus *de facto* silent because its length is off-scale and the peak does therefore not appear in the EPG.

1.2. Motivating example

As a motivating example we shall consider a case from a UK trial as reported in Gill et al. (2008), also analysed in Cowell et al. (2011):

> “An incident had occurred in a public house where the deceased had spent the evening with some friends. There was an altercation in the car park between the deceased (K_1) and several others resulting in the death of the victim. The

Table 1. Alleles, peak heights, and genotypes of individuals for an excerpt of the markers in the pub crime case.

Marker	Alleles	MC18	MC15	K_1	K_2	K_3
D2	16	189	64			16
	17	171	96			17
	22	55	0			
	23	638	507	23		
	24	673	524	24	24	
D16	11	534	256		11	11
	12	1786	1724	12	12	
	13	265	109			13
TH01	7	670	727	7		
	8	636	625	8		
	9	99	0		9	
	9.3	348	165			9.3

> alleged offenders then left the scene and went to another public house where they were seen to go into the lavatory to clean themselves."

Two blood stains, MC18 and MC15, were found at the public house lavatory. Both indicated that they were DNA mixtures of at least three individuals. The genotypes of the *defendant,* K_3, and a known individual, K_2, alleged to be present at the time of the offence, were determined together with that of the *victim* K_1 (all were males). An excerpt of the observed DNA profiles of the samples and individual genotypes are displayed in Table 1. The complete data can be found in Gill et al. (2008) and in the R-package `DNAmixtures` (Graversen, 2013).

Note that if the trace consists of DNA from exactly K_1, K_2, and K_3, the peak at allele 22 of marker D2 in MC18 would need to be due to stutter and/or spontaneous dropin, and K_2's allele 9 on marker TH01 would have dropped out of the MC15 profile. We emphasize that the entire profile would be consistent with only DNA from K_1, K_2, and K_3 being present in the traces.

1.3. Objectives of analysis

The analysis of a DNA profile can have different objectives depending on the context. The objective can be a quantification of the strength of evidence for a given hypothesis over another, or the objective may be a *deconvolution* of the profile, *i.e.* one wishes to identify likely genotypes of contributors (Perlin and Szabady, 2001; Wang et al., 2006; Tvedebrink et al., 2012a). We briefly describe the typical situations below.

Weight of evidence The available *evidence* E consists of the peak heights as observed in the EPG as well as the combined genotypes of the known individuals. It is customary to assume relevant population allele frequencies to be known. We shall return to this issue in Section 5.

To quantify the strength of the evidence against the defendant (K_3), two competing hypotheses are typically specified. One of these, usually referred to as the *prosecution hypothesis* $\mathcal{H}_p$, could in our example be that the profile has exactly three contributors who are identical to the three individuals K_1, K_2, and K_3. Since their genotypes are known, we refer to these as *known contributors.* Alternatively, $\mathcal{H}_p$ could only involve K_1 and K_3

as known contributors in addition to one or more unknown individuals who we shall name $U_1, U_2, U_3, \dots$ and refer to as *unknown contributors*. The alleles of the unknown contributors are assumed to be chosen randomly and independently from a reference population with known allele frequencies. To limit the scope of our analysis we shall here only consider scenarios which include K_1 and K_2 as contributors to the traces.

The prosecution hypothesis is then compared to what is referred to as the *defence hypothesis* $\mathcal{H}_d$, which typically would replace K_3 in $\mathcal{H}_p$ with an unknown contributor, claiming that the apparent similarity between the trace and the DNA profile of K_3 is due to chance. The defence hypothesis is not actually advocated by the defence, but is formulated strictly for comparative purposes. We emphasize that this implies that the genotypes of this unknown contributor in principle could be identical to that of K_3 although this would be extremely improbable, ignoring cases involving identical twins.

The strength of the evidence (Good, 1950; Lindley, 1977; Balding, 2005) is normally represented by the *likelihood ratio*:

$$LR = \frac{L(\mathcal{H}_p)}{L(\mathcal{H}_d)} = \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_d)}.$$

We shall follow Balding (2013) and report the *weight of evidence* as WoE $= \log_{10} LR$ in the unit *ban* introduced by Alan Turing (Good, 1979) so that one ban represents a factor 10 on the likelihood ratio. For a high quality single source DNA profile (*i.e.* from a single individual) in the SGM Plus system, values for WoE vary around 14 bans (Balding, 1999), which in most cases would be considered sufficient for establishing the identity of the individual leaving the stain. We emphasize that this unit of WoE is not absolute, but should be always be interpreted relative to a specific defence hypothesis; it could for example change — in principle in any direction — if also K_2 were replaced by an unknown individual.

The calculation of the WoE can be important both when presenting a case to the court and in the investigative phase of a trial, to decide whether it is worthwhile to search for additional independent evidence.

The numerator and denominator in the likelihood ratio are calculated based on models which shall be detailed further in Section 2 below; generally models have the form

$$\Pr(E \mid \mathcal{H}) = \sum_{\mathbf{g}} \Pr(E \mid \mathbf{g}) \Pr(\mathbf{g} \mid \mathcal{H})$$

so that the model for the conditional distribution $\Pr(E \mid \mathbf{g})$ of the evidence given the genotypes $\mathbf{g}$ of all contributors is the same for both hypotheses, whereas the hypotheses differ concerning the distribution $\Pr(\mathbf{g} \mid \mathcal{H})$ of genotypes of the contributors, as described above.

Deconvolution of DNA mixtures This calculation attempts to identify the combined genotypes across all markers of each unknown contributor to the mixture and give a list of potential genotypes of a perpetrator to use for a database search. For example, we could wish to calculate

$$\Pr\{U_1, U_2 \mid E, \mathcal{H}_{\text{inv}}\} \text{ or } \Pr\{U_1 \mid E, \mathcal{H}_{\text{inv}}\},$$

where U_1, U_2 represent genotypes of unknown contributors to the mixture under an investigative hypothesis $\mathcal{H}_{\text{inv}}$ that the trace contains DNA from three persons, one of whom is the victim K_1, or, for example, two of whom are K_1 and K_2. The calculation should be done for a range of probable combinations of genotypes of the unknown contributors.

Evidential efficiency For a single source high quality trace the WoE specializes to $-\log_{10}$ of the *match probability* i.e. the probability that a random member of the population has the specific DNA profile of K_s (Balding, 2005). We point out that the WoE against a suspect K_s based on our model can never be stronger than the WoE obtained by a single source DNA profile. To see this, let $\pi_s = \Pr(U = K_s)$ denote the match probability. Further, let $\mathcal{H}_p$ be the prosecution hypothesis involving K_s as a contributor and $\mathcal{H}_d$ the defence hypothesis, replacing the genotype K_s by that of a random individual U. We then have

$$\begin{aligned} LR = \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_d)} &= \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_d, U = K_s)\pi_s + (1-\pi_s)\Pr(E \mid \mathcal{H}_d, U \neq K_s)} \\ &= \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_p)\pi_s + (1-\pi_s)\Pr(E \mid \mathcal{H}_d, U \neq K_s)} \\ &\leq \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_p)\pi_s} = \frac{1}{\pi_s}. \end{aligned} \tag{1}$$

Thus we have WoE $= \log_{10} LR \leq -\log_{10} \pi_s$ implying that a mixed trace can never give stronger evidence than a high-quality trace from a single source. We define the *loss of evidential efficiency* $\mathrm{WL}(E \mid K_s)$ against K_s in the evidence E as

$$\mathrm{WL}(E \mid K_s) = \mathrm{WoE}(K_s)_{\max} - \mathrm{WoE}(K_s) = -\log_{10} \pi_s - \log_{10} \frac{\Pr(E \mid \mathcal{H}_p)}{\Pr(E \mid \mathcal{H}_d)}.$$

This quantity is non-negative, and indicates how many bans of WoE are lost due to the evidence being based on a mixture rather than a single source trace. Indeed, the loss of efficiency for a mixed trace can be compared to the loss induced by failure of identifying the genotype for some markers, as this increases the match probability in a similar way. For the case example in Section 1.2, using the US Caucasian allele frequencies of Butler et al. (2003), the negative logarithm of the match probability for K_3 is $-\log_{10} \pi_{K_3} = 14.5$, whereas if we ignored, say, markers D18 and D19, the weight-of-evidence would be WoE $= -\log_{10} \tilde{\pi}_{K_3} = 12.0$, losing about 2.5 bans.

In the case of deconvolution without a specified suspect, we could consider the evidence against the posterior most likely unknown person K^* under the defence hypothesis. If $\mathcal{H}_p^*$ denotes a prosecution hypothesis obtained by replacing U with K^* and $\pi^* = \Pr(U = K^*) = \Pr(U = K^* \mid \mathcal{H}_d)$ denotes the prior probability of a random individual having genotype K^*, we get

$$\begin{aligned} \mathrm{WL}(E \mid K^*) &= -\log_{10} \pi^* - \log_{10} \frac{\Pr(E \mid \mathcal{H}_p^*)}{\Pr(E \mid \mathcal{H}_d)} = -\log_{10} \frac{\Pr(E \mid \mathcal{H}_p^*)\pi^*}{\Pr(E \mid \mathcal{H}_d)} \\ &= -\log_{10} \frac{\Pr(E \mid U = K^*, \mathcal{H}_d)\Pr(U = K^* \mid \mathcal{H}_d)}{\Pr(E \mid \mathcal{H}_d)} \\ &= -\log_{10} \Pr(U = K^* \mid \mathcal{H}_d, E). \end{aligned} \tag{2}$$

Thus, this *generic loss of evidential efficiency* is equal to $-\log_{10}$ of the maximum posterior probability obtained in the deconvolution. For a single source trace that uniquely identifies the contributor, the generic loss of evidential efficiency is 0.

The loss of evidential efficiency is in particular relevant for assessing the value of DNA evidence in the investigative phase or prior to a courtroom presentation.

2. A gamma model with artefacts

In this section we present an overview of the basic model for peak heights, which is based on the gamma model described in Cowell et al. (2007a) and used in Cowell et al. (2011) with extensions describing artefacts. It should be emphasized that the model used here is different from the latter in two main respects: firstly it uses absolute peak heights instead of relative peak heights which enables direct treatment of dropout by thresholding; secondly, it has a different description of stutter. As a consequence, computations are greatly simplified so analysis is possible for a high number of unknown potential contributors.

2.1. Basic model

We consider I potential contributors to a DNA mixture. Let there be M markers used in the analysis of the mixture with marker m having A_m allelic types, $m = 1, \dots, M$. Let ϕ_i denote the *fraction* of DNA from individual i prior to PCR amplification, with $\phi = (\phi_1, \phi_2, \dots, \phi_I)$ denoting the fractions from all contributors. Thus $\phi_i \geq 0$ and $\sum_{i=1}^{I} \phi_i = 1$. For identifiability of the parameters we further assume that the fractions ϕ_i for unknown contributors are listed in non-increasing order, *i.e.* $\phi_a \geq \phi_b$ if $a < b$ are both unknown individuals. It is assumed that these pre-amplification fractions of DNA are constant across markers.

For a specific marker m and allele a, the model is describing the total *peak height* H_a. Ignoring artefacts for the moment, the model makes the following further assumptions: Each contribution H_{ia} from an individual i to the peak height at allele a has a gamma distribution, $H_{ia} \sim \Gamma(\rho\phi_i n_{ia}, \eta)$, where $\Gamma(\alpha, \beta)$ denotes the distribution with density

$$f(h \mid \alpha, \beta) = \frac{h^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-h/\beta} \text{ for } h > 0.$$

For $\alpha = 0$, $\Gamma(0, \beta)$ is the distribution degenerate at 0. Here ρ is proportional to the total amount of DNA in the mixture prior to amplification; the number of alleles of type a carried by individual i is denoted by n_{ia}; and the parameter η determines the scale. We note that the individual contributions H_{ia} are unobservable. Indeed, when artefacts are added to the model, the peak heights $H_a = \sum_i H_{ia}$ themselves are unobserved; they will be modified by stutter and dropout as described below in Section 2.2.

For notational simplicity we have suppressed the dependence of the heights H_{ia}^m and other quantities on markers m. It is important for the model that ϕ_i are the same for all markers, whereas one would typically expect other parameters to be marker dependent and possibly dependent on fragment length in case of degraded DNA (Tvedebrink et al., 2012b). We note that the assumption of a common scale parameter η^m for the individual contributions H_{ia}^m for different alleles a enables a simple interpretation of the model in terms of adding independent contributions to peaks both from proper alleles and from stutter, see below.

It becomes practical to introduce the *effective number* $B_a(\phi, \mathbf{n})$ *of alleles of type* a where $B_a(\phi, \mathbf{n}) = \sum_i \phi_i n_{ia}$, and $\mathbf{n} = (n_{ia}, i = 1, \dots I; a = 1, \dots, A)$; we note that as $\sum_i \phi_i = 1$, we have that $\sum_a B_a(\phi, \mathbf{n})$ is constant over markers and equal to two in the diploid case. The term indicates that the peak at allele a behaves as it would for a single contributor with $B_a(\phi, \mathbf{n})$ alleles of type a. If the contributor fraction ϕ is interpreted as a probability distribution, it is the expected number of alleles of type a.

As a sum of independent gamma distributed contributions, the peak height H_a at allele a is gamma distributed as

$$H_a \sim \Gamma\{\rho B_a(\phi, \mathbf{n}), \eta\}. \tag{3}$$

The distribution of H_a has expectation $\rho\eta B_a(\phi, \mathbf{n})$ and variance $\rho\eta^2 B_a(\phi, \mathbf{n})$. Thus the mean of H_a is approximately proportional to the amount of DNA of type a and the variance is proportional to the mean. We let $\mu = \rho\eta$ and $\sigma = 1/\sqrt{\rho}$. Then — in a trace with only one heterozygous diploid contributor and no artefacts — μ is the *mean peak height* and σ the *coefficient of variation* for peak heights; hence σ becomes a measure of generic peak imbalance. In the presentation of results we shall often use (σ, μ) instead of (ρ, η) because of their more direct interpretability.

2.2. Incorporating artefacts

The model described above does not incorporate a number of important artefacts as described in Section 1.1. This section is devoted to the necessary modifications needed to take these into account.

Stutter We represent stutter by decomposing the individual contributions H_{ia} to peak heights in the EPG into H^s_{ia} and H^0_{ia} so that

$$H_{ia} = H^s_{ia} + H^0_{ia}$$

where H^s_{ia} represents the DNA originating from individual i of allelic type a that stutters to $a-1$ by losing a repeat number, and H^0_{ia} represents the remainder that goes through the PCR process undamaged. We let the components be independent and gamma distributed as

$$H^s_{ia} \sim \Gamma\{\rho\xi\phi_i n_{ia}, \eta\}, \quad H^0_{ia} \sim \Gamma\{\rho(1-\xi)\phi_i n_{ia}, \eta\},$$

where ξ is the *mean stutter proportion.* The total peak height observed at allele a is then

$$H_a = \sum_i H^0_{ia} + \sum_i H^s_{i,a+1} = H^0_a + H^s_{a+1}.$$

Note that now H_a has a slightly different meaning than in (3) above as it has potentially received stutter from the allele above and lost stutter to the allele below. We ignore the possibility that H_a might have tiny stutter contributions $H^s_{a+2}, H^s_{a+3}, \cdots$ from peaks at $a+2, a+3, \ldots$. Including this is straightforward in principle, but it has strong effects on the complexity of computations (Graversen and Lauritzen, 2013a). The new peak height H_a is also gamma distributed as

$$H_a \sim \Gamma\{\rho D_a(\phi, \xi, \mathbf{n}), \eta\}$$

where now

$$D_a(\phi, \xi, \mathbf{n}) = (1-\xi)B_a(\phi, \mathbf{n}) + \xi B_{a+1}(\phi, \mathbf{n}) \tag{4}$$

are the *effective allele counts after stutter.* Here and elsewhere we have let $B_{A+1}(\phi, \mathbf{n}) = 0$. The relative contribution lost to stutter

$$X_a = \frac{H^s_a}{H^s_a + H^0_a}$$

thus follows a beta distribution $\mathcal{B}\{\xi\rho B_a(\phi,\mathbf{n}),(1-\xi)\rho B_a(\phi,\mathbf{n})\}$ with mean $\mathrm{E}(X_a)=\xi$.

Notice that our use of stutter proportion X_a differs from standard practice which uses the ratio between the stutter peak and the parent peak (Butler, 2005). A peak in stutter position may itself include contributions from proper alleles, and the parent peak may itself include stutter contributions from an allele with higher repeat number. Even in the simple event that a peak is entirely due to stutter from a parent peak, which has not itself been inflated by stutter, we have $H_a = H_a^0$ and $H_a^s = H_{a-1}$ and thus our $X_a = H_{a-1}/(H_{a-1}+H_a)$ rather than the conventional H_{a-1}/H_a.

Dropout The next step in our model is concerned with the fact that in mixtures some alleles are not observed; either because no peak can be identified, or because the peak is below some detection threshold. This phenomenon is prominent for small amounts of DNA. Here we take the consequence of the model and represent dropout by a total peak height below the chosen threshold C, as described in Section 1.1. In other words, we do not observe the height H_a as just described, but rather Z_a, where

$$Z_a = \begin{cases} H_a & \text{if } H_a \geq C \\ 0 & \text{otherwise.} \end{cases}$$

This implies that the probability that a specific allele is not observed is

$$P(Z_a = 0 \mid \mathbf{n}) = G\{C; \rho D_a(\phi,\xi,\mathbf{n}),\eta\},$$

where G denotes the cumulative distribution function of the gamma distribution.

Allelic dropout for single source traces was studied by Tvedebrink et al. (2009, 2012) who fitted a logistic regression model to the dropout probability as

$$P(Z_a = 0 \mid \bar{h}) = \frac{\alpha\bar{h}^\beta}{1+\alpha\bar{h}^\beta} \tag{5}$$

where $\bar{h}$ is an average of observed peak heights above threshold calculated across all markers. They found that one could use the same value of $\beta = -4.35$ for all markers, whereas α was marker dependent. The independent variable $\bar{h}$ was used as a proxy for total amount of DNA.

In our model, the theoretical mean peak height $\mu = \rho\eta$ would be a similar proxy for the total amount of DNA. To compare the two models, we assume that $\bar{h}$ approximately corresponds to the mean peak height μ. Thus our model has

$$P(Z_a = 0 \mid \mu) = G\{C; \mu/\eta, \eta\}, \tag{6}$$

which can then be compared with (5) where $\bar{h}$ is replaced by μ. A selection of curves (6) for $C = 50$ and values of η corresponding to the maximum likelihood estimate and upper and lower 99% confidence limits in our case example, see Section 4; together with curves (5) for $\beta = -4.35$ and representative values of α, are superimposed in Fig. 2. We note that the dropout model based on the gamma distribution tends to have lower dropout rates than the logistic model for small amounts of DNA.

Both dropout models implicitly define a relation between the dropout probability d for a single allele and the corresponding dropout probability D for a homozygote. Balding and Buckleton (2009) argue that these should satisfy $D < d^2$. This relation is satisfied by any

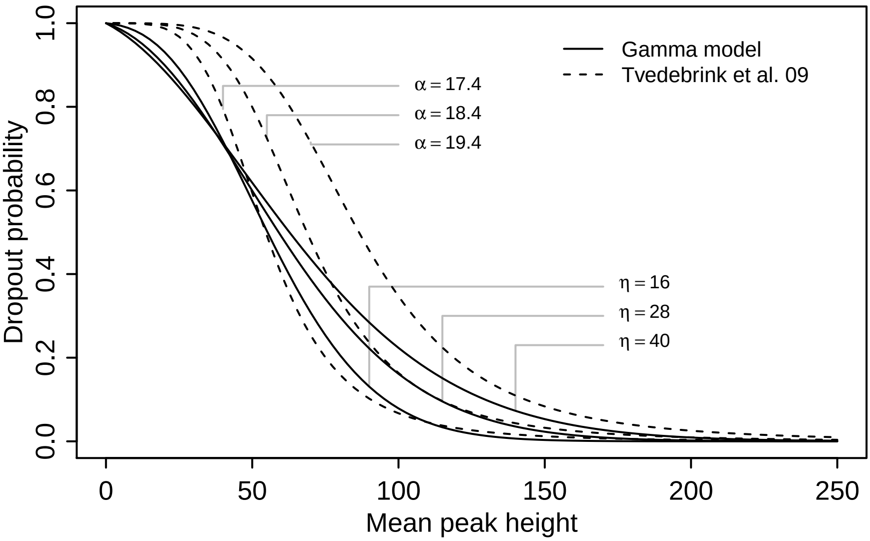

Fig. 2. The probability of dropout of a single allele as a function of mean peak height. The curves with full lines correspond to our gamma model whereas the dashed curves correspond to the logistic model.

threshold model — hence also for our gamma model — since if peak height contributions Y_1 and Y_2 from each allele are independent and identically distributed, we have

$$D = P(Y_1 + Y_2 < C) < P(Y_1 < C)^2 = d^2.$$

For the model (5) we have

$$\frac{d}{1-d} = \alpha \bar{h}^{\beta}; \qquad \frac{D}{1-D} = \alpha(2\bar{h})^{\beta} = 2^{\beta}\frac{d}{1-d}$$

and thus

$$D = \frac{2^{\beta}d}{1+(2^{\beta}-1)d} \tag{7}$$

independently of the value of α. The curve in (7) is displayed in Fig. 3 for $\beta = -4.35$ together with the upper bound $D = d^2$ and a selection of curves from the gamma model with $C = 50$ and the same values for η as in Fig. 2. We note that although (7) actually just crosses the curve $D = d^2$ for very small values of d, this has hardly any practical significance, as also pointed out in Tvedebrink et al. (2012).

Other artefacts We have chosen to represent spontaneous dropin of alleles by additional unknown contributors with very small amounts of DNA, implying that most of their alleles will drop out with high probability. With the computational methods described in Graversen and Lauritzen (2013a) and Section 2.4, this representation is amply feasible. Also, the possibility of a silent allele is easily incorporated into the model, simply by adding an allele that never results in an observed peak, see Section 4.3 for details. Note that we have

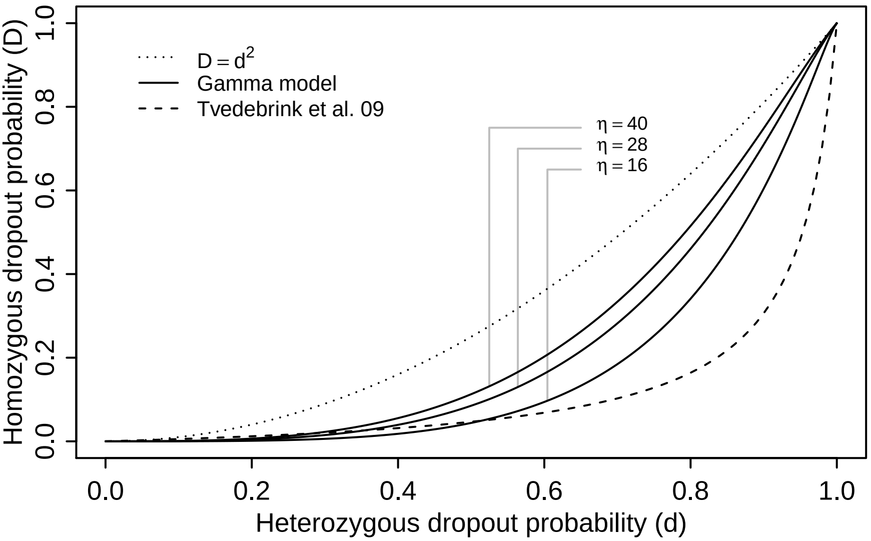

Fig. 3. The probability of homozygous dropout D as a function of the dropout probability of a single allele for the gamma model and for the logistic model of Tvedebrink et al. (2009). The solid curves correspond to the gamma model for different values of η.

modelled dropout entirely as a threshold phenomenon, ignoring the fact that apparatus failure could be an alternative explanation.

2.3. Joint likelihood function

For given genotypes of the contributors $\mathbf{n}$, given proportions ϕ, and given values of the parameters (ρ, ξ, η), all observed peak heights are independent. Thus, the conditional likelihood function based on the observations $\mathbf{z} = \{z_{ma}, m = 1, \dots, M; a = 1, \dots, A_m\}$ for all markers m and alleles a is

$$L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) = \prod_m \prod_a L_{ma}(z_{ma})$$

where

$$L_{ma}(z_{ma}) = \begin{cases} g\{z_{ma}; \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{if } z_{ma} \geq C \\ G\{C; \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{otherwise,} \end{cases}$$

with g and G denoting the gamma density and cumulative distribution function respectively and D_a are the effective allele counts after stutter in (4). For simplicity of notation, we have suppressed potential marker and allele dependence of ρ, η, and ξ in the above formulae.

For a given hypothesis $\mathcal{H}$, the full likelihood is obtained by summing over all possible combinations of genotypes $\mathbf{n}$ with probabilities $P(\mathbf{n} \mid \mathcal{H})$ associated with the hypothesis to give

$$L(\mathcal{H}) = \Pr(E \mid \mathcal{H}) = \sum_{\mathbf{n}} L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) P(\mathbf{n} \mid \mathcal{H}). \tag{8}$$

The number of terms in this sum is huge for a hypothesis which involves several unknown contributors to the mixture, but can be calculated by Bayesian network techniques (Graversen and Lauritzen, 2013a); some details are given in Section 2.4 below.

Estimating unknown parameters The likelihood function $L(\mathcal{H})$ involves a number of parameters (ρ, ξ, ϕ, η) which may be completely or partially unknown.

One way of dealing with this is to make the likelihood as large as possible for each of the competing hypotheses and thus calculate

$$\hat{L}(\mathcal{H}) = \sup_{\rho,\xi,\phi,\eta} \sum_{\mathbf{n}} L(\rho, \xi, \phi, \eta \,|\, \mathbf{z}, \mathbf{n}) P(\mathbf{n} \,|\, \mathcal{H}),$$

corresponding to using maximum likelihood estimates for the unknown parameters, but different estimates under the competing hypothesis. Since the likelihood function (8) can be efficiently computed, it is feasible to maximize it using appropriate numerical optimization methods.

Using maximized likelihood ratios preserves the property that the WoE against a suspect K_s in a mixed trace based on our model can never be stronger than what is obtained by a matching single source DNA profile. To see this, recall that $\pi_s = \Pr(U = K_s)$ denotes the match probability, *i.e.* the probability that a random individual has the genotype of the suspect K_s. As before, let $\mathcal{H}_p$ be the prosecution hypothesis, involving K_s as a contributor and $\mathcal{H}_d$ be the defence hypothesis, replacing K_s by a random individual U. Let further $\hat{\psi}_p = (\hat{\rho}_p, \hat{\xi}_p, \hat{\phi}_p, \hat{\eta}_p)$ be the maximum likelihood estimates of the unknown parameters under the prosecution hypothesis and $\hat{\psi}_d$ the estimates under the defence hypothesis. We then have

$$\begin{aligned}
\mathrm{WoE}(K_s \,|\, \hat{\psi}_d, \hat{\psi}_d) &= \log_{10} \frac{\Pr(E \,|\, \mathcal{H}_p, \hat{\psi}_d)}{\Pr(E \,|\, \mathcal{H}_d, \hat{\psi}_d)} \leq \log_{10} \frac{\Pr(E \,|\, \mathcal{H}_p, \hat{\psi}_p)}{\Pr(E \,|\, \mathcal{H}_d, \hat{\psi}_d)} \\
&= \mathrm{WoE}(K_s \,|\, \hat{\psi}_p, \hat{\psi}_d) \leq \log_{10} \frac{\Pr(E \,|\, \mathcal{H}_p, \hat{\psi}_p)}{\Pr(E \,|\, \mathcal{H}_d, \hat{\psi}_p)} \\
&= \mathrm{WoE}(K_s \,|\, \hat{\psi}_p, \hat{\psi}_p) \leq -\log_{10} \pi_s = \mathrm{WoE}(K_s)_{\max},
\end{aligned}$$

where the last inequality is obtained as in (1). Thus, also when parameters are estimated, a mixed trace can never give stronger evidence than a high-quality trace from a single source.

2.4. Computation

As mentioned in Section 2.3, the main computational challenge is associated with the marginalisation over all unknown DNA profiles needed for evaluating the likelihood function. This is particularly challenging when the number of contributors increases, as the number of terms in (8) becomes intractably huge so brute force summation is infeasible. To enable efficient computation of the sum, we use an appropriate Bayesian network to represent the genotypes of the unknown contributors and the observations. As markers are independent for fixed model parameters, the model can be represented by a separate Bayesian network for each marker. Here we shall only briefly describe the basic structure of these networks and refer the reader to Graversen and Lauritzen (2013a) for further details on the networks and associated computational issues.

Representation of genotypes As customary we assume a reference population in Hardy-Weinberg equilibrium so we can consider the two alleles of an individual chosen at random with allele frequencies $(q_1, \ldots, q_A)$. We recall from Section 2.2 that the distribution of peak height at allele a depends on both n_{ia} and $n_{i,a+1}$ for both known and unknown contributors i. To enable simple computation of the terms in the likelihood function (8), we therefore represent the genotype for contributor i by a vector of allele counts $(n_{i1}, \ldots, n_{iA})$. These vectors then follow independent multinomial distributions with $\sum_a n_{ia} = 2$.

Using properties of the multinomial distribution we can describe this distribution sequentially as follows. The number n_{i1} of the first allelic type follows a binomial distribution $\text{Bin}(2, q_1)$. Denote by $S_{ia} = \sum_{b \leq a} n_{ib}$ the number of alleles of type 1 to a that contributor i possesses. For any subsequent allelic type $a+1$, it holds that, conditionally on S_{ia}, the number of alleles of type $a+1$ is independent of the specific allocations $(n_{i1}, \ldots, n_{ia})$ already made. Furthermore, $n_{i,a+1}$ is also binomially distributed with $n_{i,a+1} | S_{ia} \sim \text{Bin}(2 - S_{ia}, q_{a+1} / \sum_{b \geq a+1} q_b)$. Thus, adding the partial sums of allele counts to the network allows the genotype to be modelled by a Markov structure as displayed in Fig. 4.

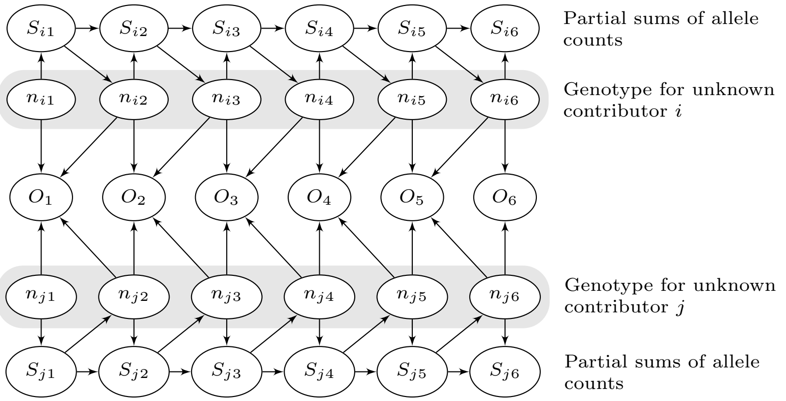

Fig. 4. Bayesian network modelling the genotypes of 2 unknown contributors i and j and peak height observations for a marker with 6 possible allelic types.

Including peak height information The information about peak height observations is incorporated via binary nodes O_a in Fig. 4; these represent whether a peak for allele a has been seen. By suitable specification of the conditional distribution of O_a given its parent nodes, conditioning on O_a being TRUE or FALSE will correspond to conditioning on the observed peak heights Z_a; this enables in particular fast evaluation of the likelihood function (8) by a single propagation in the Bayesian network (Graversen and Lauritzen, 2013a).

The computational effort is exponential in the number of unknown contributors. However, the Markov representation of the genotypes themselves ensures the complexity to grow linearly with the number of allelic types A, rather than polynomially as in previously used representations (Cowell et al., 2007a,b, 2011).

Actual computation time clearly depends on a number of factors. The computations for the analysis in Section 4 were performed in a script available as supplementary material to this article. Running this script took 18 minutes on a standard desktop with the current version of DNAmixtures, RHugin, and HUGIN. The computation time increases with the number of unknown contributors; the maximization of the likelihood function for five unknown contributors used around three hours on the same machine. Maximization of the likelihood function for six unknown contributors as used in Fig. 8 was done on a machine with more memory. See Graversen and Lauritzen (2013a) for further discussion of the computational issues.

3. Related literature

Methods for the interpretation of DNA profiles arising from mixtures can be classified into:

a) those mostly based on qualitative information, *i.e.* information about allele presence or absence in the DNA mixture;

b) those that also use quantitative information by taking into account both allele presence and peak intensities.

We will first discuss some recent relevant papers that base their model prevalently on allele presence and secondly we describe those that also exploit the information in the peak intensities. A list of publications relating to DNA mixtures is maintained at the website of the National Institute of Standards and Technology (NIST)†.

3.1. Methods based on qualitative information

Recently, the International Society for Forensic Genetics published recommendations based on discrete models for forensic analysis of single source DNA profiles (Gill et al., 2012). They consider a single locus (DNA marker) and at that locus allow for artefacts such as dropin and dropout, including the probability of their occurrence in the likelihood ratio computations. Under the independence assumptions of this model, they state that it readily extends to multiple loci and DNA mixtures. Gill et al. (2012) is largely based on Curran et al. (2005) and Balding and Buckleton (2009).

Curran et al. (2005) give a set based method for likelihood ratio calculations which includes multiple contributors and the analysis of replicate runs. This was an improvement on previous guidelines for the analysis of LTDNA profiles where alleles were not reported unless they were duplicated in replicate runs (Gill et al., 2000). They consider the probability of contamination and dropout, but not of stutter. The model is implemented in Gill et al. (2007). For the interpretation of LTDNA, Gill et al. (2008) include the probabilities of dropout, contamination and stutter; peaks in stutter positions are considered ambiguous alleles and included in the calculation. Furthermore, they account for dropin by increasing the number of potential contributors to the mixture. They also study the robustness of the WoE to misleading evidence in favour of the prosecution.

Balding and Buckleton (2009) consider a discrete model for interpreting LTDNA mixtures where peaks are classified as present, absent, or masking. The set of masking alleles is defined as every peak above a designated threshold that either corresponds to an allele of a known contributor, is in a stutter position to that allele or to another peak of sufficient height. Based on an extension of this paper (Balding, 2013), Balding wrote a suite of R

†http://www.cstl.nist.gov/strbase/mixture.htm

functions likeLTD‡ built for a range of crime scene DNA profiles, involving complex mixtures, uncertain allele designations, dropin and dropout, degradation, stutter, relatedness of alternative possible contributors, as well as replicated runs. Alleles are defined as present, absent, or uncertain; dropout is modelled using Tvedebrink et al. (2009, 2012). Haned et al. (2011) estimate dropout probabilities using different logistic models for haploid and diploid cells in combination with a simulation model of the PCR process (Gill et al., 2005). Similarly, Haned et al. (2012) and Haned and Gill (2011) adopt a model based on allele presence that analyses the sensitivity to dropin and dropout parameters. The papers mentioned in this paragraph do not make direct use of peak heights, but peak heights enter implicitly in the process of allele classification.

3.2. Methods based on quantitative information

Early attempts to exploit peak intensity information include Evett et al. (1998); Perlin and Szabady (2001); Wang et al. (2006); Bill et al. (2005); and Curran (2008). Cowell et al. (2007a) introduced the use of gamma distributions to model peak area variability observed in the PCR amplification of DNA in mixtures, and presented a normal-approximation version of the model in Cowell et al. (2007b). Both of these papers carry out calculations using exact propagation algorithms for Bayesian networks, but require discretization of a parameter representing the relative amounts of DNA from each contributor to the mixture. Both evidential and deconvolution analyses were presented. The suitability of the gamma distribution was investigated in Cowell (2009), using the simulation model of Gill et al. (2005).

The gamma model of Cowell et al. (2007a) was extended in Cowell et al. (2011) to handle stutter, dropout and silent alleles. Stutter was represented in Bayesian networks by means of a discretization to allow probability propagation. The authors showed how the Bayesian network representation readily enabled the simultaneous analysis of two DNA mixtures that are thought to have contributors in common. The authors noted that the computational complexity of the model increased significantly with the total number of contributors, limiting the practical use of their methods to at most four contributors.

Perlin et al. (2011) present a model for DNA mixture interpretation based on a normal model extending the method in Perlin and Szabady (2001) in a direction similar to Cowell et al. (2007b). They use a fully Bayesian approach where all parameters are given a prior distribution. The model is used both for identification of contributors to a DNA mixture and for deconvolution. The model does not incorporate stutter, and dropout is accounted for by a background variance parameter. The performance of the model is reported for 16 two-person mixture samples, however the performance for mixtures of DNA from more than two people is not reported.

Puch-Solis et al. (2013) is based on using the gamma distribution for peak heights as in Cowell et al. (2007a, 2011). Since DNA degrades more the greater the DNA fragment length (Tvedebrink et al., 2012b), they use a correction for this phenomenon. Their methodology is different to the one presented here in that: i) they exploit the common scale parameter for peak heights to derive Dirichlet distributions for relative peak heights; ii) they use a preprocessing step for classifying peaks as possible stutter peaks; iii) they use a more complex dropout model; iv) they discretize the parameter representing the proportion of DNA from each contributor and use a finer grid on extreme intervals to better capture unbalanced

‡https://baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code

Table 2. Maximum likelihood estimates based on MC15 and MC18 analysed separately.

Prosecution hypothesis					Defence hypothesis				
	MC15		MC18			MC15		MC18	
Parameter	Est.	SE	Est.	SE	Parameter	Est.	SE	Est.	SE
μ	913	35	1056	39	μ	914	41	1056	41
σ	0.171	0.018	0.166	0.017	σ	0.198	0.023	0.172	0.019
ξ	0.074	0.014	0.085	0.016	ξ	0.072	0.019	0.085	0.017
ϕ_{K_1}	0.821	0.020	0.706	0.022	ϕ_{K_1}	0.798	0.028	0.698	0.026
ϕ_{K_2}	0.047	0.014	0.091	0.016	ϕ_{K_2}	0.039	0.020	0.096	0.018
ϕ_{K_3}	0.124	0.015	0.194	0.018	ϕ_{U_1}	0.081	0.013	0.193	0.020
ϕ_U	0.008	0.019	0.009	0.018	ϕ_{U_2}	0.081	0.013	0.013	0.021
$\log_{10}\hat{L}(\mathcal{H}_p)$	-118.0		-130.1		$\log_{10}\hat{L}(\mathcal{H}_d)$	-129.3		-143.4	

mixtures; v) all alleles not corresponding to peaks above threshold are lumped together into a single compound allele. They discuss identification of contributors to the mixture and not deconvolution. The paper presents examples of two person mixtures and does not analyze replicate runs or multiple traces.

Tvedebrink et al. (2010) evaluate the weight of evidence for two person mixtures, using a multivariate normal distribution of peak heights. Based on DNA mixtures from controlled experiments they find a linear relationship between peak height and area and between means and variances of peak height measurements. Controlled experiments are also used for estimating the probability of allelic dropout per locus using logistic regression (Tvedebrink et al., 2009, 2012).

Recently, Taylor et al. (2013) used a log-normal model for the ratio between observed and expected peak heights.

4. Case analysis

We proceed to illustrate the methodology developed in Section 2 by applying it to the case described in Section 1.2. We have used a threshold of $C = 50$ for both traces MC15 and MC18; note that for MC18 the peaks for alleles 21 and 25 at marker FGA are of height 49 and 39 and thus below the threshold. Using a threshold of $C = 39$ for this profile makes a negligible difference to the results.

The computations were all performed with the R-package `DNAmixtures` (Graversen, 2013) which is available from `dnamixtures.r-forge.r-project.org`. The R-package interfaces the HUGIN API (Hugin Expert A/S, 2012) through the R-package `RHugin` (Konis, 2012). The likelihood functions are maximized numerically using `Rsolnp` (Ghalanos and Theussl, 2012; Ye, 1987). Approximate standard errors of estimates are based on the inverse Hessian of the likelihood function found by numerical derivation using `numDeriv` (Gilbert and Varadhan, 2012). The population allele frequencies used were taken from the US Caucasian database in Butler et al. (2003).

4.1. Weight of evidence

Analysis of MC15 The maximum likelihood estimates and their approximate standard errors under the prosecution hypothesis $\mathcal{H}_p : K_1 \& K_2 \& K_3 \& U$ and defence hypothesis $\mathcal{H}_d : K_1 \& K_2 \& U_1 \& U_2$ are given in Table 2. The last unknown contributor is mainly included to allow for potential spontaneous dropin. We note that the estimates for the parameters (μ, σ, ξ) are very similar under the two hypotheses, the defence hypothesis suggesting a slightly larger coefficient of variation, both a little less than 20%, i.e. indicating large variability of the peak heights. The mean stutter proportion ξ is estimated to just under 7.5%. The estimates of the contributor fractions ϕ under the prosecution hypothesis agree well with the estimates $(0.80, 0.06, 0.12, 0.04)$ found in Table 3 of Gill et al. (2008). The resulting WoE is $\log_{10}(LR) = -118.0 - (-129.3) = 11.3$ bans. For the prosecution hypothesis we can ignore spontaneous dropin, the standard deviation of ϕ_U indicating this may well be zero. Under the defence hypothesis there are difficulties distinguishing the two unknown contributors in that their contributor fractions are estimated to the same value. Since the estimates for ϕ_{U_i} are on the boundary of the parameter space, their standard errors are given for the model which further restricts them to be equal as $\phi_{U_1} = \phi_{U_2}$. Removing one unknown contributor for both hypotheses increases the WoE to 12.1 bans. This should be compared to the analysis in Cowell et al. (2011) which for hypotheses with three contributors gave a WoE around 10 bans. It should also be compared to the upper bound of 14.5 bans calculated from the match probability for K_3.

Thus our model gives considerably stronger WoE than found previously, although much less than would have been possible for a perfect single source trace. The loss of efficiency compared to a single source trace is similar to the loss from leaving out, for example, markers D18 and D19; see Section 1.3.

Analysis of MC18 The maximum likelihood estimates and their approximate standard errors under the defence and prosecution hypotheses are given in Table 2.

We note again that the estimates for (μ, σ, ξ) are very similar under the two hypotheses, the defence hypothesis suggesting a slightly larger coefficient of variation, but the same order of magnitude as for MC15. Also here the estimates of the contributor fractions ϕ under the prosecution hypothesis agree well with the estimates $(0.67, 0.11, 0.19, 0.04)$ in Table 2 of Gill et al. (2008). In contrast to MC15, the defence hypothesis identifies an unbalanced contribution to MC18 from the unknown contributors, the major unknown contributor representing a fraction similar to that of the defendant K_3 under the prosecution hypothesis. The estimates of the mean stutter proportions are slightly larger than found for MC15, although the standard errors suggest that the differences are well within what could be expected from random variation. The resulting WoE for this trace becomes 13.3 bans, which compares to losing the information in marker D18, say.

Standard deviations for contributor fractions indicate that one unknown contributor can be removed for both hypotheses; this has no essential effect on the WoE which remains at 13.3 bans. We compare again to Cowell et al. (2011) which gave a WoE around 10 bans, again a weaker evidential value than what we obtain from our model.

Combined analysis of MC15 and MC18 The combined analysis of more than one DNA profile can be very informative when computing the WoE but especially for deconvolution of mixed traces, see for example Section 4.2. This is particularly true when there might be high dropout probabilities in one or more of the profiles. Multiple profiles with the same contributors can, for example, be obtained from different parts of a crime stain, but also from different stains found at the same scene. In addition, as we shall see, the combined analysis may considerably improve the precision of parameter estimates.

For a combined analysis of the two traces there is a range of possibilities. The simplest possible combined analysis assumes both for the prosecution and defence that the unknown

Table 3. Maximum likelihood estimates when information in the traces MC15 and MC18 is combined.

	Prosecution hypothesis					Defence hypothesis			
	MC15		MC18			MC15		MC18	
Par.	Est.	SE	Est.	SE	Par.	Est.	SE	Est.	SE
μ	914	36	1055	38	μ	914	36	1055	39
σ	0.175	0.013	0.163	0.012	σ	0.178	0.013	0.165	0.013
ξ	0.079	0.011	0.079	0.011	ξ	0.079	0.011	0.079	0.011
ϕ_{K_1}	0.822	0.020	0.705	0.021	ϕ_{K_1}	0.820	0.021	0.702	0.023
ϕ_{K_2}	0.048	0.014	0.090	0.016	ϕ_{K_2}	0.049	0.014	0.091	0.017
ϕ_{K_3}	0.125	0.016	0.193	0.018	ϕ_{U_1}	0.123	0.016	0.193	0.018
ϕ_U	0.006	0.018	0.012	0.017	ϕ_{U_2}	0.008	0.019	0.014	0.018
$\log_{10} \hat{L}(\mathcal{H}_p)$				-248.2	$\log_{10} \hat{L}(\mathcal{H}_d)$				-262.3

contributors to the two traces are different and unrelated. Under this assumption, the traces are independent and the joint likelihood ratio is obtained by multiplying the two likelihood ratios found above to yield overwhelmingly strong evidence at 24.6 bans. However, this defence hypothesis is extremely unfavourable for the defence; as we shall see below, the observed evidence is much more probable under the assumption that the unknown persons are identical for the two traces, yielding more sensible numbers for the WoE. This highlights the relativity of the WoE to the hypotheses chosen: a very high WoE could in principle be due to the choice of a completely inadequate defence hypothesis.

As the traces are dependent when unknown contributors are shared, a fresh calculation is needed to find the likelihood ratio. In this calculation we assume that the two traces have identical mean stutter proportions ξ and scale parameters η. This seems reasonable since these parameters refer to variability in the determination of peak heights rather than to the specific traces. In addition we assume that the unknown contributors are the same for the two profiles, yielding the results displayed in Table 3.

We note that the standard errors in Table 3 are not dissimilar to previous values, apart from the standard errors of $\hat{\sigma}$ and $\hat{\xi}$, which are reduced considerably by combining information from the two traces. The estimates for most parameters are strikingly similar under the two hypotheses. The fraction attributed to the defendant K_3 under the prosecution hypothesis is attributed to the major unknown contributor U_1 under the defence hypothesis. The interpretations under the prosecution and defence hypotheses only disagree on the identity of this contributor. The resulting WoE for the combination of traces becomes 14.1 bans. This evidence has strength similar to that of a single source trace with a perfect match to the defendant K_3 which would yield a WoE of 14.5 bans. Thus the loss of evidential efficiency $\mathrm{WL}(E \mid K_3)$ against K_3 is about 0.4 bans. As for the single trace analyses one unknown contributor is redundant and could be removed from the analysis, leaving the WoE unchanged at 14.1 bans. Cowell et al. (2011) gave a WoE of 8.9 bans using their model.

Above we have analysed the combined traces under the assumption that the unknown contributors to the two traces are the same. To assess whether this assumption is reasonable, we can in principle compare the maximized likelihoods to those allowing possibly different unknown contributors under the two hypotheses. Due to the many one-sided constraints on the parameters, p-values for such tests are not easily calculated although the techniques in El Barmi and Dykstra (1999) should be applicable. In the present analyses we believe there is no reason to assume that the contributors are different.

Table 4. Most probable genotypes of the unknown contributor U under the hypothesis K_1 & K_2 & U when information in the traces MC15 and MC18 is combined.

	Marker			
Rank	D16	D18	D19	Profile probability
1	$(11, 13)$	$(12, 16)$	$(14, 15)$	0.436
2	$(11, 13)$	$(12, 14)$	$(14, 15)$	0.129
3	$(11, 13)$	$(12, 16)$	$(13, 15)$	0.070
4	$(11, 13)$	$(12, 16)$	$(13, 14)$	0.052
5	$(12, 13)$	$(12, 16)$	$(14, 15)$	0.047
Total				0.734

We also note from Table 3 that the mixture proportions ϕ appear to be similar for the two traces, although the standard deviations indicate that they are not identical; indeed a likelihood ratio test for their identity is rejected under both hypotheses with p-values around 0.0003.

For curiosity we also analysed MC15 and MC18 using `likeLTD` obtaining a WoE of 8.8 bans. This number is not directly comparable to our analysis above as it refers to a slightly different database of allele frequencies and a coancestry parameter $F_{ST} = 0.02$ is applied together with a sampling adjustment, resulting in a larger match probability for K_3 and thus a smaller maximal WoE of 13.2 bans. Furthermore, `likeLTD` assumes that the contributor fractions are identical for the two traces since they are treated as replicate runs of the same trace.

To compare with the analysis obtained by `likeLTD` we have made an analysis with our model assuming equal contributor fractions, using the same database, sampling adjustment, and coancestry parameter. This correction can readily be made as it only involves appropriate modification of the allele frequencies (Balding, 2013). Using a likelihood function based only on observed peak presence or absence, plugging in parameter estimates based on the observed peak heights, we get a WoE of 10.0 bans. The corresponding WoE based on observed peak heights is 12.7 bans. It seems fair to conclude that properly exploiting the information in peak heights gives a more efficient analysis of the traces than what can be obtained from allele classifications alone; this could be important for the analysis of mixtures more complex than those analysed here.

4.2. Mixture deconvolution

For mixture deconvolution we consider the traces jointly under the defence hypothesis modified by removing one unknown contributor and identify the five most probable full genotypes of the unknown individual U. For most markers, these genotypes share that of the defendant K_3; indeed the only variations are on markers D16, D18, and D19, where the most probable configurations are displayed in Table 4. The markers not represented in the table share the genotype of the defendant for all five most probable combinations. We note that the profile with the top rank is that of the defendant, and the evidence gives a probability to the unknown person having precisely this profile of 44%. The probability that the true profile of the unknown contributor is among the five genotypes listed is 73%.

If we make a deconvolution of the single traces, the identification of the genotype of U is less clear. The distribution of the unknown profile becomes much more diffuse reflected by a drastic reduction in the probability of the most probable profiles: in the case of MC15 to 0.2% and for MC18 to 11%. The probability that the true profile of the unknown contributor is among the five most probable profiles is reduced to 1% and 39% compared to 73% for the traces combined. Also, for the single trace deconvolutions, the profile of the defendant is ranked fourth for both traces rather than first. Thus it is apparent that combining trace information for deconvolution purposes can be of considerable value as it gives a much clearer explanation of the samples.

4.3. Interpreting artefacts

Our model does not impose at the outset that a specific peak or allele is due to stutter, or has dropped out, or is silent. One of the features of `DNAmixtures` (Graversen, 2013) is to produce Bayesian networks for each marker with peak height evidence propagated. Thanks to the flexibility of Bayesian networks we can then elaborate these so as to explicitly represent artefacts.

We can also make simple modifications of the model to allow for the possible presence of silent alleles. In this way, we can answer queries like: what is the probability that an observed peak is due only to stutter? that a specific allele has dropped out? that an unknown contributor possesses a silent allele?

Identifying stutter and dropout To enable a precise analysis we shall say that an observed peak at a is *due to stutter* if no contributors possess allele a. Similarly we say that an allele a has *dropped out* if at least one contributor possesses allele a, but no peak has been observed above threshold, *i.e.* $Z_a = 0$.

To investigate such events we introduce variables Y_a which explicitly represent the presence or absence of the allele a as follows

$$Y_a = \begin{cases} 1 & \text{if } \sum_i n_{ia} > 0 \\ 0 & \text{if } \sum_i n_{ia} = 0. \end{cases}$$

These variables can readily be included in the network as nodes with parents n_{ia} and probability propagation would then yield $P(Y_a = 1 \mid \mathbf{z})$ and $P(Y_a = 0 \mid \mathbf{z})$. For an allele a where a peak has been observed, the probability that the peak is due to stutter alone is now $P(\text{stutter} \mid \mathbf{z}) = P(Y_a = 0 \mid \mathbf{z})$, whereas if no peak has been observed at a, the probability that the allele a has dropped out is $P(\text{dropout} \mid \mathbf{z}) = P(Y_a = 1 \mid \mathbf{z})$. A pre-classification of some alleles as, say, stutter or dropout corresponds to conditioning on specific values of Y_a; the consequences of this conditioning can be obtained by a single probability propagation. A revised mixture analysis can then be performed conditionally on the pre-classification, should this be desired. We note that this form for preprocessing of data would still be consistent with our general model specification.

We now revisit the defence hypothesis $\mathcal{H}_d : K_1 \mathbin{\&} K_2 \mathbin{\&} U_1 \mathbin{\&} U_2$ using the parameter estimates in Table 3 and obtain the results in Table 5. Conforming with the remarks in Section 1.2 we note that the peak at allele 22 for D2 has high probability (93%) of being a stutter peak. Similarly, we note that the allele 9 for marker TH01 has certainly dropped out, consistent with the fact that K_2 possesses this allele. We also note, for example, that it is rather probable (33%) that allele 6 for marker TH01 is actually present in the mixture despite the fact that neither of K_1 and K_2 have that allele represented in their genotype and no peak has been observed. This is partly due to this allele having a high frequency (23%) in the reference population; but, since no peak has been observed, the fraction of DNA from of an unknown contributor possessing this allele is likely to be very small.

Table 5. Posterior probabilities for stutter and dropout.

Trace	Marker	Allele a	z_a	$P(\text{stutter}\,\vert\,\mathbf{z})$	$P(\text{dropout}\,\vert\,\mathbf{z})$
MC18	D2	15	0		0.003
		16	189	0	
		17	171	0	
		18	0		0.156
		19	0		0.219
		20	0		0.274
		21	0		0.083
		22	55	0.927	
		23	638	0	
		24	673	0	
		25	0		0.178
		26	0		0.060
		27	0		0.003
MC15	TH01	5	0		0.004
		6	0		0.328
		7	727	0	
		8	625	0	
		9	0		1.000
		9.3	165	0	
		10	0		0.018
		11	0		0.004

Unsurprisingly, the combined analysis of two traces gives more informative results on the analysis of artefacts than what emerges from separate analyses of the two traces. We refrain from reporting the separate analyses here.

Silent alleles The possibility that the unknown contributors might have a silent allele is incorporated in the model by adding an extra allelic type 0 (silent) to the genotype representations as displayed in Fig. 5. We now let $n_{i0} \sim \text{Bin}(2, q_0)$, $S_{i0} = n_{i0}$, $S_{i1} = n_{i1} + S_{i0}$, and further $n_{i1} \,|\, S_{i0} \sim \text{Bin}(2 - S_{i0}, q_1)$, where q_0 is the probability of an allele being silent and q_1 indicates the database allele frequency. Since no peak can be observed for the silent allele there is no observational node O_0. This corresponds to a modification of the basic model where the allele frequencies $q_a, a > 0$ are interpreted as the relative

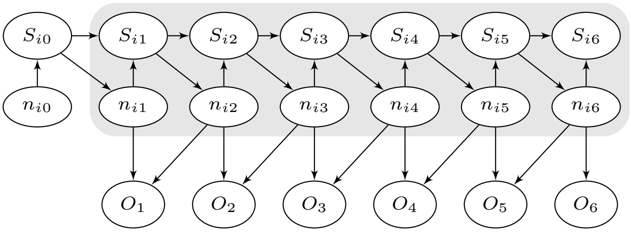

Fig. 5. Modified network allowing for silent alleles. The nodes in the shaded area are those in the original genotype representation in Fig. 4.

Table 6. Excerpt of a DNA profile with peak heights of a single individual.

Marker	Allele	Height
D8	13	4364
D21	28	2646
	29	2490
CSF1PO	11	2695
D3	14	2249
	16	2205
TH01	6	1268
	9.3	1294
TPOX	8	1394
	11	1526
AMEL	X	4289
D5	11	2053
	13	1827
FGA	23	2444

frequencies of non-silent alleles, so the probability that a random allele is of type a is equal to $(1 - q_0)q_a$ if $a > 0$.

4.4. Single source analysis

Although our model is developed for the analysis of mixtures, it can also be useful in the analysis of a DNA profile from a single source; in particular as the peak heights can be informative about the presence of silent alleles. What appears to be a homozygous genotype at some marker may not be so; an alternative explanation is that we see only one allele of a heterozygous genotype, the other allele being silent. We would expect this to be reflected in a peak height that is smaller than expected and it will clearly affect the evidential interpretation of DNA profiles.

Table 6 shows an excerpt of a DNA profile of a single individual J, having apparent homozygous genotypes on markers D8, CSF1PO and FGA. Note that the peak heights are rather large, as is customary in single source data with a large amount of DNA.

Assuming that the DNA profile in Table 6 came from two unknown contributors and setting the threshold $C = 50$, we obtain estimates for the fraction of DNA from each unknown contributor of $(\hat{\phi}_1, \hat{\phi}_2) = (1, 0)$ which clearly indicate that the DNA trace comes from a single individual. The estimates of the other parameters are $\hat{\mu} = 1806$, $\hat{\sigma} = 0.290$, $\hat{\xi} = 0$. The combination of a high mean peak height $\hat{\mu} = 1806$ and a vanishing stutter proportion points to the data having been preprocessed so that peaks classified in the laboratory as stutter have been removed. If we instead set the threshold at $C = 500$ to ensure that the removed stutter peaks are well below threshold and assume that the data are from a single individual, we get the estimates $\hat{\mu} = 1854$, $\hat{\sigma} = 0.278$ and $\hat{\xi} = 0.038$. Clearly, there is very little information about the stutter proportion when stutter peaks have been removed, resulting in a very flat profile likelihood for the stutter proportion as shown in Fig. 6; indeed a 95% confidence interval for ξ would be ranging from 0 to 11.8%.

In Fig. 7 we see the posterior probabilities that J has a silent allele for the markers showing apparent homozygosity: D8, CSF1PO, and FGA. As might be expected, for D8 where the peak height is relatively high, the probability of a silent allele is almost zero. Markers

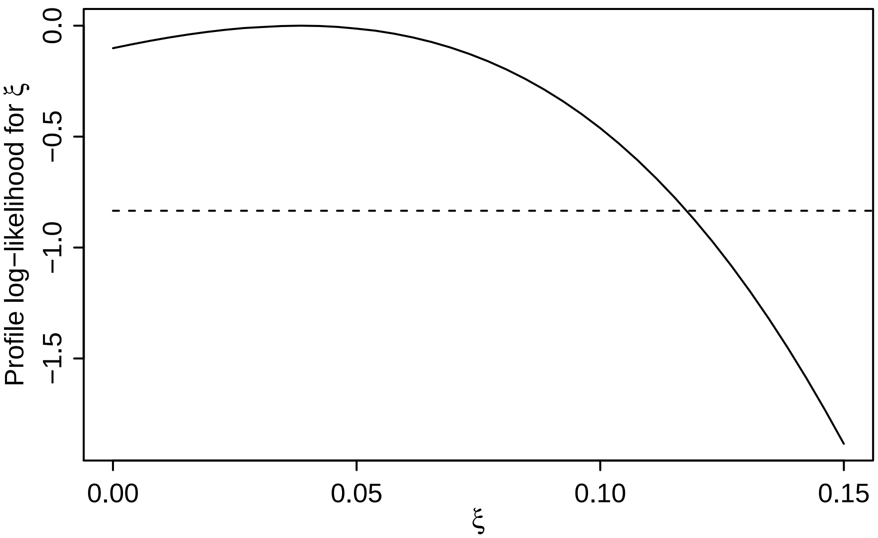

Fig. 6. Profile log-likelihood $\max_{\mu,\sigma} \log_{10} L(\mu, \sigma, \xi) - \log_{10} L(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ for ξ for the single source trace thresholded at $C = 500$. Values of ξ with log-likelihood above the horizontal line constitute an approximate 95% confidence interval.

CSF1PO and FGA have peak heights around 60% and 56% that of D8, and correspondingly higher probabilities of J possessing a silent allele.

However, care should be taken with interpreting this result as FGA and CSF1PO often have a lower peak height than other markers, being at the end of the dye panel and thus corresponding to a longer fragment length, see also Section 5.2 below. Evidence calculations and deconvolutions may be less sensitive to this fact, but it could be of crucial importance for the identification of silent alleles.

5. Discussion

On the previous pages we have described a simple and versatile model for description of mixed DNA profiles and demonstrated how it could be applied to case analysis. We should emphasize that all our computations are direct consequences of our model and exact apart from the use of deterministic numerical optimization methods. Clearly, the model itself represents an approximation to reality, as any model will, but we believe it is a virtue that any further analysis — whether this be computation of likelihood ratios, deconvolution, or mixture interpretation — does not need additional approximations, modifications of the model, nor any *ad hoc* heuristics. Also, our method does not rely on a subjective pre-processing of the DNA profile with a manual or automatic system which identifies which alleles have been observed; such a pre-processing step constitutes an additional potential source of error.

In addition to the case discussed in this article, we have tried our methodology on a number of real challenging cases, including difficult LTDNA mixtures, with very sensible and robust results and believe that the methodology is now sufficiently developed for use in

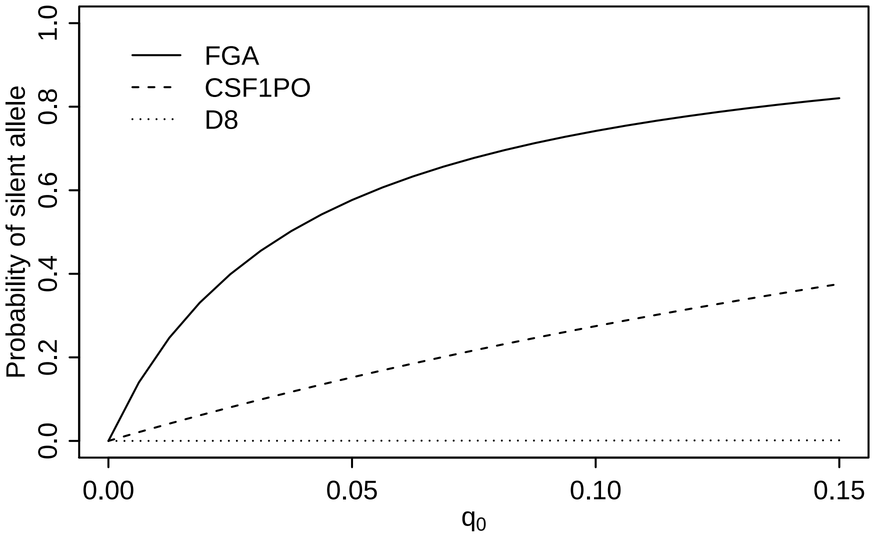

Fig. 7. Posterior probabilities that the individual has a silent allele as a function of the apriori probability q_0 of having a silent allele.

proper case work. However, for any methodology there is room for improvement and below we shall briefly discuss some issues worthy of further attention.

We note in passing that the model could potentially also be useful in disputed paternity cases or other types of pedigree analysis where the DNA of some of the actors is available only in small quantities as in LTDNA, or is degraded.

5.1. Number of contributors

One issue is associated with determining the number of unknown contributors to the sample. The samples considered need at least three unknown contributors to be well explained. We have in the previous analyses introduced an additional unknown contributor to allow for spontaneous dropin, but showed that such an additional contributor had most likely only contributed tiny amounts of DNA to the mixture, if anything at all. Typically, a heuristic argument is used where additional contributors are not added unless they appeared to be necessary for explaining the DNA profile in question. However, Buckleton et al. (2007) and Biedermann et al. (2012) point out the risks in using this kind of approach.

As we operate with maximized likelihoods under hypotheses that there is at most a fixed number k of contributors, it is obvious that the maximized likelihood will increase — or at least not decrease — with the number of unknown contributors, both for the defence and prosecution. This follows from the fact that the hypothesis of at most k contributors is a sub-hypothesis of that with $k + 1$ contributors, so we are maximizing over a larger space when increasing k. The phenomenon is illustrated in Fig. 8, which shows the maximized likelihoods and corresponding ratios for up to eight contributors, *i.e.* six unknown contributors for the defence hypothesis. The likelihoods and ratios are normalized relative to what was obtained for three contributors. The likelihood does indeed increase and it increases more for the defence than for the prosecution, so the likelihood

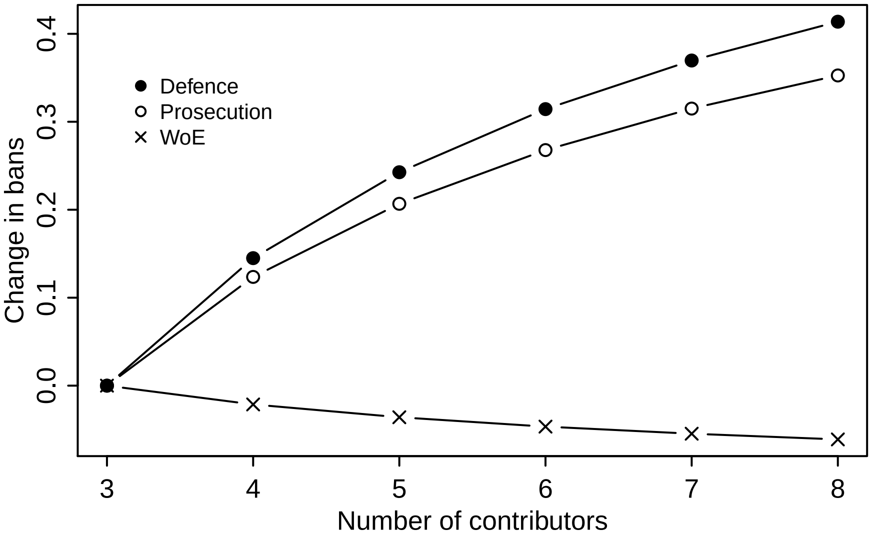

Fig. 8. The maximized log-likelihood functions based on combining MC15 and MC18 for varying number of contributors and hypotheses, relative to their value for three contributors and displayed in bans.

ratio decreases with the number of contributors. The change is tiny: for eight contributors the WoE has decreased to 14 bans from 14.1 bans for three contributors, hardly of any significance. Even if the defence is allowed eight contributors and the prosecution stays with three, we get WoE = 13.7, which is still close to what would be obtained for a perfect match on nine markers and in most cases would imply identification of K_3 as a contributor beyond reasonable doubt.

Inspection of the maximum likelihood estimates reveals that the five unknown contributors associated with the smallest amount of DNA have between them supplied less than 3.7% of the total amount of DNA shared in equal proportions, *i.e.* each about 0.75% of the total amount of DNA. Bearing in mind that the total solution of DNA contains about 100 human cells, this means that effectively all of their alleles (which are not shared with the three major contributors or in stutter position to those) have not been observed. The presence of these unknown contributors just serves the purpose of explaining minor peak imbalances and seems irrelevant for the substantial interpretation of the mixture.

We believe that the phenomenon of ever increasing likelihoods is an artefact of using direct maximum likelihood without a penalty for complexity for estimating the unknown parameters. There are various ways of overcoming this problem: for example, one could use a convention that unknown contributors are only added if their presence is statistically significant at some fixed level; one could maximize a likelihood that is penalized for having many contributors; or one could postulate a uniform prior distribution for the unknown fractions ϕ and use an integrated rather than maximized likelihood. All of these would be acceptable but will be essentially *ad hoc* so would need to be established by an agreed convention. We point out that the methodology for bounding the number of contributors in Lauritzen and Mortera (2002) would not be applicable as this strongly exploits the assumption that no allele has dropped out; the problem in our case is that one could have an infinite number of contributors with so little DNA that their alleles have only been observed through minor perturbations of the peak heights.

5.2. Extensions and modifications

The analyses presented in this paper have been made with the simplest possible variant of the model with sensible and rather robust results. However, there are some obvious potential improvements to be made.

Marker dependence of parameters We have assumed that the parameters (μ, σ, ξ) are identical for all markers. This clearly contradicts findings in the literature (Butler, 2005; Puch-Solis et al., 2013) which indicate that both mean peak heights and stutter proportions depend on marker, fragment length, and may also be different for different dye panels in the EPG. Variation in these parameters can easily be incorporated in the model. However, as this could drastically increase the number of parameters, it would be best to do so by using prior information from laboratory experiments, for example by scaling peak heights as indicated in Puch-Solis et al. (2013) or by adding prior information by penalizing the likelihood function to ensure that parameter estimates are near those expected.

Degradation For strongly degraded DNA samples it is necessary to adjust the model by correcting for fragment length. From the analysis in Tvedebrink et al. (2012b) it seems that the simplest and most natural way to do so would be to assume the parameter ρ in the gamma distribution — being proportional to the amount of DNA — to depend exponentially on the fragment length λ as

$$\log \rho = \delta\lambda + \zeta.$$

This would just add one additional parameter δ representing the level of degradation of DNA to the model and would not in itself create specific difficulties for the methodology, although the complexity of the maximum likelihood estimation clearly increases whenever new parameters are added.

Dropout We have chosen a simple approach to modelling dropout by taking the consequence of our gamma model and using the probabilities for being below thresholds as dropout probabilities. The dropout model of Tvedebrink et al. (2009, 2012) can be seen as using a logistic distribution rather than a gamma distribution for this purpose. The choice of the logistic model seems primarily motivated by the ability to analyse dropout data using standard software and may not in itself have a firm theoretical foundation. In any case it would be interesting to see whether any of the gamma and logistic models would best describe the experimental data behind Tvedebrink et al. (2009). We note that since Tvedebrink et al. (2009) uses average peak height as a proxy for the amount of DNA and the relation between peak height and amount may be different in LTDNA analysis, their parameter estimates may not be directly suitable for this purpose but need modification, as described in Tvedebrink et al. (2012). Our estimation process would naturally fit the parameters to any specific sample in question.

5.3. Other issues

Model criticism An important point that we have not touched upon in the previous analyses is the need to validate the model used for any given dataset. We find that it is not enough to compare the likelihoods for two competing hypothesis if neither of them can be demonstrated to give a plausible explanation of the data at hand. Promising graphical methods for systematic model validation and criticism based on its internal predictions and residual analysis are important and under development (Graversen and Lauritzen, 2013a). These methods exploit that the computational structure of a Bayesian network naturally enables fast prediction of peak heights based on partial information from the profiles. An example is given in Fig. 9 which shows the conditional probability transform of peak heights for MC15 and MC18 under the prosecution hypothesis against quantiles of the uniform distribution, indicating an excellent fit of the gamma model.

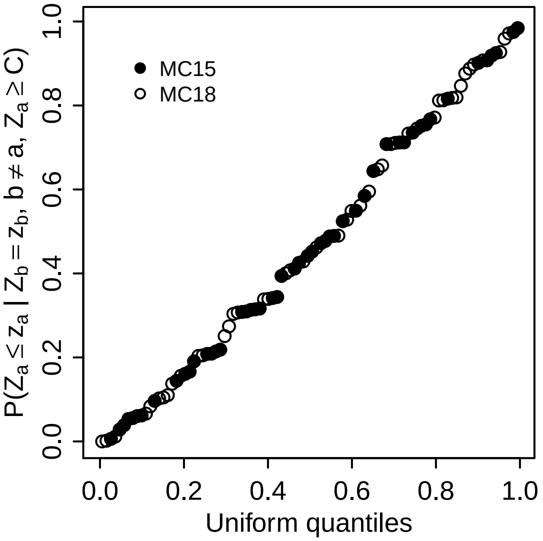

Fig. 9. Conditional probability transform of observed peak heights for MC15 and MC18 vs. uniform quantiles.

Uncertain allele frequencies We have in the previous analysis treated the allele frequencies as fixed and known, although in principle they also are parameters which should be estimated. Using allele counts from databases directly as frequencies corresponds to estimating allele frequencies by maximum likelihood and using the estimated frequencies as if they were known. This conforms with the way other parameters are treated in the present paper, but ignores the inherent uncertainty associated with the frequencies, see also Curran et al. (2002) and Curran (2005) for a discussion of this aspect. It would be valuable to develop a scheme similar to that in Green and Mortera (2009) for incorporating both this uncertainty and kinship corrections based on the possibility of alleles that are *identical by descent* (IBD) and we hope to address this in future work. As noted in Green and Mortera (2009), the simple kinship correction used in Balding (2013) and in our case analysis ignores the dependence between markers created by IBD or by population heterogeneity, which may generally still lead to an overstatement of the WoE.

Estimation uncertainty We note that in addition to the sampling uncertainty associated with the allele frequencies, we also have statistical uncertainty associated with estimation of other model parameters. Although we have indicated the uncertainty of the latter by giving approximate standard errors based on the Hessian of the log-likelihood function, we have neither made a detailed analysis that justifies this, nor have we incorporated this additional uncertainty into our analysis. However it would be feasible to do so along the lines described in Graversen and Lauritzen (2013b) and this will be further developed in the future.

Acknowledgements

This research was partially supported by Università Roma Tre "PESFoG Fondi Internazionalizzazione" Grant. We are indebted to Marina Dobosz for providing the EPGs in Fig. 1 and Francesca Scarnicci for providing the data for Section 4.4. In addition we have benefited from Ate Kloosterman's explanation on details of the PCR process, and discussions with other participants of the PESFoG workshop in "Villino Volterra" Ariccia 2012.

References

Aitken, C. C. G. and F. Taroni (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, UK: John Wliey and Sons Ltd.

Balding, D. (2013). Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences of the United States of America 110*(30), 12241–12246.

Balding, D. J. (1999). When can a DNA profile be regarded as unique. *Science and Justice 39*(4), 257–260.

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Statistics in Practice. Chichester, UK: Wiley.

Balding, D. J. and J. Buckleton (2009). Interpreting low template DNA profiles. *Forensic Science International: Genetics 4*(1), 1 – 10.

Biedermann, A., S. Bozza, K. Konis, and F. Taroni (2012). Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method. *Forensic Science International: Genetics 6*(6), 689 – 696.

Bill, M., P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, and J. Buckleton (2005). PENDULUM – a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International 148*(2-3), 181–189.

Buckleton, J. S., J. M. Curran, and P. Gill (2007). Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics 1*(1), 20–28.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Second Edition.* Elsevier Academic Press.

Butler, J. M., R. Schoske, P. M. Vallone, J. W. Redman, and M. C. Kline (2003). Allele frequencies for 15 autosomal loci on U.S. Caucasian, African American, and Hispanic populations. *Journal of Forensic Science 48*(4), 908–911. Available online at www.astm.org.

Cowell, R. G. (2009). Validation of an STR peak area model. *Forensic Science International: Genetics 3*(3), 193 – 199.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007a). A gamma model for DNA mixture analyses. *Bayesian Analysis 2*(2), 333–348.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International 166*(1), 28–34.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2011). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics 5*(3), 202–209.

Curran, J., P. Gill, and M. Bill (2005). Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Science International 148*(1), 47 – 53.

Curran, J. M. (2005). An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, Probability, and Risk 4*, 115–126.

Curran, J. M. (2008). A MCMC method for resolving two person mixtures. *Science and Justice 48*, 168–177.

Curran, J. M., J. S. Buckelton, C. M. Triggs, and B. S. Weir (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science and Justice 42*(1), 29–37.

El Barmi, H. and R. L. Dykstra (1999). Likelihood ratio test against a set of inequality constraints. *Journal of Nonparametric Statistics 11*(1-3), 233–250.

Evett, I., P. Gill, and J. Lambert (1998). Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Science 43*(1), 62–69.

Findlay, I., R. Frazier, A. Taylor, and A. Urquhart (1997). Single cell DNA fingerprinting for forensic applications. *Nature 389*, 555–556.

Ghalanos, A. and S. Theussl (2012). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method.* R package version 1.12.

Gilbert, P. and R. Varadhan (2012). *numDeriv: Accurate Numerical Derivatives.* R package version 2012.9-1.

Gill, P., J. Curran, and K. Elliot (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research 33*(2), 632–643.

Gill, P., J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, and J. Lambert (2008). Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics 2*(2), 91–103.

Gill, P., L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, and B. Weir (2012). DNA commission of the international society of forensic genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International: Genetics 6*(6), 679 – 688.

Gill, P., A. Kirkham, and J. Curran (2007). LoComatioN: a software tool for the analysis of low copy number DNA profiles. *Forensic Science International 166*(2–3), 128 – 138.

Gill, P., J. Whitaker, C. Flaxman, N. Brown, and J. Buckleton (2000). An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International 112*(1), 17–40.

Good, I. J. (1950). *Probability and the Weighing of Evidence.* C. Griffin London.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in world war II. *Biometrika 66*(2), 393–396.

Graversen, T. (2013). *DNAmixtures: Statistical Inference for Mixed Traces of DNA.* R package version 0.1-0, `dnamixtures.r-forge.r-project.org/`.

Graversen, T. and S. Lauritzen (2013a). Computational aspects of DNA mixture analysis. arXiv:1302.4956.

Graversen, T. and S. Lauritzen (2013b). Estimation of parameters in DNA mixture analysis. *Journal of Applied Statistics 40*(11), 2423–2436.

Green, P. J. and J. Mortera (2009). Sensitivity of inferences in forensic genetics to assumptions about founder genes. *Annals of Applied Statistics 3*(2), 731–763.

Haned, H., T. Egeland, D. Pontier, L. Pène, and P. Gill (2011). Estimating drop-out probabilities in forensic DNA samples: A simulation approach to evaluate different models. *Forensic Science International: Genetics 5*(5), 525 – 531.

Haned, H. and P. Gill (2011). Analysis of complex DNA mixtures using the Forensim package. *Forensic Science International: Genetics Supplement Series 3*(1), 79 – 80.

Haned, H., K. Slooten, and P. Gill (2012). Exploratory data analysis for the interpretation of low template DNA mixtures. *Forensic Science International: Genetics 6*(6), 762 – 774.

Hugin Expert A/S (2012). *Hugin API Reference Manual, Version 7.6.* Aalborg, Denmark: Hugin Expert A/S.

Jeffreys, A. J., A. J. Wilson, and S. L. Thein (1985). Individual specific 'fingerprints' of human DNA. *Nature 316*, 76–79.

Konis, K. (2012). *RHugin.* R package version 7.6-2.

Lauritzen, S. L. and J. Mortera (2002). Bounding the number of contributors to mixed DNA stains. *Forensic Science International 130*, 125–126.

Lindley, D. (1977). A problem in forensic science. *Biometrika 64*(2), 207–213.

Perlin, M. and B. Szabady (2001). Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Science 46*, 1372–1378.

Perlin, M. W., M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, and B. W. Duceman (2011). Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences 56*(6), 1430–1447.

Puch-Solis, R., L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, and D. Balding (2013). Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics 7*(5), 555–563.

Stolovitzky, G. and G. Cecchi (1996). Efficiency of DNA replication in the polymerase chain reaction. *Proceedings of the National Academy of Sciences 93*, 12947–12952.

Sun, F. (1995). The polymerase chain reaction and branching processes. *Journal of Computational Biology 2*(1), 63–86.

Taylor, D., J.-A. Bright, and J. Buckleton (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics 7*(5), 516 – 528.

Tvedebrink, T., P. S. Eriksen, M. Asplund, H. S. Mogensen, and N. Morling (2012). Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation. *Forensic Science International: Genetics 6*(2), 263 – 267.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222 – 226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. *Applied Statistics 59*(5), 855 – 874.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2012a). Identifying contributors of DNA mixtures by means of quantitative information of STR typing. *Journal of Computational Biology 19*(7), 887–902.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2012b). Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Science International: Genetics 6*(1), 97 – 101.

Wang, T., N. Xue, and J. D. Birdwell (2006). Least-square deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of Forensic Sciences 51*(6), 1284–1297.

Ye, Y. (1987). *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming.* Ph. D. thesis, Department of Electrical Engineering, Stanford University.